Applied and Environmental Microbiology, June 2010, p. 3886–3897 Vol. 76, No. 12
0099-2240/10/$12.00 doi:10.1128/AEM.02953-09

Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes[▽][†]

Anna Y. Pei,[1]‡ William E. Oberdorf,[2]‡ Carlos W. Nossa,[2] Ankush Agarwal,[2] Pooja Chokshi,[5] Erika A. Gerz,[2] Zhida Jin,[2] Peng Lee,[3] Liying Yang,[3] Michael Poles,[2] Stuart M. Brown,[4] Steven Sotero,[4] Todd DeSantis,[7] Eoin Brodie,[7] Karen Nelson,[8] and Zhiheng Pei[2,3,6]*

New York University College of Arts and Science, New York, New York 10012[1]; Departments of Medicine[2] and Pathology[3] and Center for Health Informatics and Bioinformatics,[4] New York University School of Medicine, New York, New York 10016; Tufts University College of Arts and Sciences, Medford, Massachusetts 02155[5]; Department of Veterans Affairs New York Harbor Healthcare System, New York, New York 10010[6]; Ecology Department, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720[7]; and J. Craig Venter Institute, Rockville, Maryland 20850[8]

Received 7 December 2009/Accepted 19 April 2010

Analysis of intragenomic variation of 16S rRNA genes is a unique approach to examining the concept of ribosomal constraints on rRNA genes; the degree of variation is an important parameter to consider for estimation of the diversity of a complex microbiome in the recently initiated Human Microbiome Project (http://nihroadmap.nih.gov/hmp). The current GenBank database has a collection of 883 prokaryotic genomes representing 568 unique species, of which 425 species contained 2 to 15 copies of 16S rRNA genes per genome (2.22 ± 0.81). Sequence diversity among the 16S rRNA genes in a genome was found in 235 species (from 0.06% to 20.38%; 0.55% ± 1.46%). Compared with the 16S rRNA-based threshold for operational definition of species (1 to 1.3% diversity), the diversity was borderline (between 1% and 1.3%) in 10 species and >1.3% in 14 species. The diversified 16S rRNA genes in *Haloarcula marismortui* (diversity, 5.63%) and *Thermoanaerobacter tengcongensis* (6.70%) were highly conserved at the 2° structure level, while the diversified gene in *B. afzelii* (20.38%) appears to be a pseudogene. The diversified genes in the remaining 21 species were also conserved, except for a truncated 16S rRNA gene in "*Candidatus* Protochlamydia amoebophila." Thus, this survey of intragenomic diversity of 16S rRNA genes provides strong evidence supporting the theory of ribosomal constraint. Taxonomic classification using the 16S rRNA-based operational threshold could misclassify a number of species into more than one species, leading to an overestimation of the diversity of a complex microbiome. This phenomenon is especially seen in 7 bacterial species associated with the human microbiome or diseases.

rRNA genes are widely used for estimation of evolutionary history and taxonomic assignment of individual organisms (14, 26, 50–52). The choice of rRNA genes as optimal tools for such purposes is based on both observations and assumptions of ribosomal conservation (13, 50). rRNA genes are essential components of the ribosome, which consists of >50 proteins and three classes of RNA molecules; precise spatial relationships may be essential for assembly of functional ribosomes, constraining rRNA genes from drastic change (9, 13). In bacteria, the three rRNA genes are organized into a gene cluster which is expressed as single operon, which may be present in multiple copies in the genome. In organisms with multiple rRNA gene operons, the gene sequences tend to evolve in concert. It is generally believed that copies of rRNA genes within an organism are subject to a homogenization process through homologous recombination, also known as gene conversion (18), a form of concerted evolution that maintains their fit within the ribosome. The homogenization process may involve short domains without affecting the entire sequence of each gene (8).

However, significant differences between copies of rRNA genes in single organisms, albeit few, have been discovered in all three domains of life and in all three classes of rRNA genes. The amphibian *Xenopus laevis* and the loach *Misgurnus fossilis* have two types of 5S rRNA genes that are specific to either somatic or oocyte ribosomes (30, 48). The parasite *Plasmodium berghei* contains two types of 18S rRNA genes that differ at 3.5% of the nucleotide positions and are life cycle stage specific (17). The metazoan *Dugesia mediterranea* possesses two types of 18S rRNA genes with 8% dissimilarity (6). The archaeon *Haloarcula marismortui* contains two distinct types of 16S rRNA genes that differ by 5% (32, 33). In the domain *Bacteria*, the actinomycete *Thermobispora bispora* contains two types of 16S rRNA genes that differ by 6.4% (47). Copies of the 16S rRNA genes and 23S rRNA genes of the actinomycete *Thermospora chromogena* differ by approximately 6 and 10%, respectively (54). Paralogous copies of rRNA genes with different sequences may have functionally distinct roles.

Divergent evolution between rRNA genes in the same genome may corrupt the record of evolutionary history and obscure the true identity of an organism. Substantial variation, if it occurs, may lead to the artificial classification of an organism into more than one species. For a cultivable organism, this problem can be resolved by cloning rRNA genes from a pure

[*] Corresponding author. Mailing address: Department of Medicine, New York University School of Medicine, New York, NY 10016. Phone: (212) 951-5492. Fax: (212) 263-4108. E-mail: zhiheng.pei@med.nyu.edu.

‡ A. Y. Pei and W. E. Oberdorf contributed equally to this work.

† Supplemental material for this article may be found at http://aem.asm.org/.

[▽] Published ahead of print on 23 April 2010.

culture of the organism to identify the degree of variation. However, most environmental surveys and the recently initiated Human Microbiome Project (HMP) (http://nihroadmap.nih.gov/hmp/) (34) use cultivation-independent techniques to examine microbiomes that contain mixed species. In the case of the HMP, it is hoped that this approach may identify some idiopathic diseases that are caused by alterations in the microbiome in humans. In this type of study, it may be impossible to trace all rRNA genes observed back to their original host. For example, in the phylum TM7, multiple 16S rRNA gene sequences have been reported (21), but it is not known whether they belong to multiple species or to the same bacterium with a high degree of intragenomic variation among rRNA gene paralogs. Due to the limited number of microorganisms for which nucleotide sequences are available for all copies of the rRNA genes, intragenomic variation among 16S rRNA genes, and the likelihood of pyrosequencing errors (25, 40), the potential to overestimate the diversity of a microbiome exists.

Coenye et al. analyzed 55 bacterial genomes and found the intragenomic heterogeneity between multiple 16S rRNA genes in these genomes was below the common threshold (1 to 1.3%) for distinguishing species (44) and was unlikely to have a profound effect on the classification of taxa (10). The analysis of 76 whole genomes by Acinas et al. revealed the extreme diversity (11.6%) of 16S rRNA genes in *Thermoanaerobacter tengcongensis* (2). These early analyses of intragenomic variation of 16S rRNA genes were limited to a small number of available whole genomes. With the increasing number of whole microbial genomes available from the National Center for Biotechnology Information (NCBI), the extent of diversity among the paralogous 16S rRNA genes within single organisms can now be more thoroughly assessed. In the present study, we (i) addressed the theory of 16S rRNA conservation by systematic evaluation of intragenomic diversity of 16S rRNA sequences in completely sequenced prokaryotic genomes to assess its effect on the accuracy of 16S rRNA-based molecular taxonomy and (ii) examined whether previously observed ribosomal constraints on conservation of 2° structures are uniformly applicable at the intragenomic level.

MATERIALS AND METHODS

Annotation of rRNA genes. Gene sequences were obtained from the Complete Microbial Genomes database at the NCBI website (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). For some species, more than one genome was available. To avoid overrepresentation of any species, only the most completely annotated genome was selected to represent a species.

Analysis of intragenomic diversity in 16S rRNA genes. Genomes that contained only a single 16S rRNA gene were not further analyzed. Copies of 16S rRNA genes from each remaining genome were aligned with Sequencher (Gene Codes Corporation, Ann Arbor, MI). To calculate diversity, the number of revealed mismatches and insertions was divided by the total number of positions, including gaps in the alignment. If gaps were determined to be caused by intervening sequences (IVS) (inserts of >10 bp), they were recorded and removed, and sequences were realigned and reanalyzed. To determine the effect of masking hypervariable positions, the two 16S rRNA genes with the largest difference in a genome were aligned using NAST (12) in the Greengenes 7,682-character format. The aligned sequences were compared to the Lane mask (27) to mask away all but 1,287 conserved columns (lanes) of aligned characters. After common gaps had been removed, diversity between the two sequences was calculated.

Comparison of 2° structures. Secondary structures were analyzed based on minimizing free energy, using RNAstructure (31) and visualized by Rnaviz (11), with experimentally defined 16S rRNA or the consensus 16S rRNA models (53) used for reference. To compare two related 2° structures, a mismatch was defined as conserved if located in a loop or located in a stem, but causing GC:GU conversions or covariation resulting in no change in base pairing. In contrast, a nonconserved mismatch that alters base pairing and structure was classified as a stem-loop transition. When base pairing and/or structure was significantly altered and did not allow for more detailed comparison, localized segments including 50 bp upstream and 50 bp downstream of the area of diversity were analyzed at various energy levels. The two most similar structures at the lowest energy levels were compared. If no identical structures were found, these areas were classified into one of two groups. In areas of regional complex rearrangement (37), the bases pair differently and result in a slight change to the sizes and relative location of 2° structures in the immediate area, but the structures look similar topographically. In regional alteration of 2° structures, highly concentrated areas of substitutions result in topographically different structures. Both of these could not be further broken down into covariation, GU = GC, indels, or in-loop changes, and were considered on the whole to be localized, nonconservative changes. For 16S rRNA genes that displayed high levels of regional diversity, the regions in question were also folded using the KnetFold program (5). This folding method creates secondary structures based on multiple sequences. The output from KnetFold was entered into jViz.RNA 2.0 in order to visualize the secondary structure (49). jViz.RNA 2.0 allows for the creation of complex 2° structures that may contain pseudoknots. The multiple-sequence folding was verified using another program named Murlet (23).

For *T. tengcongensis* and *Borrelia afzelii*, mismatches also were classified by the position-specific relative variability rate, calculated from the consensus 16S rRNA gene model based on an alignment of 3,407 bacterial 16S rRNA genes (53). Positions were classified as variable or nonvariable according to the substitution rate relative to the average substitution rate of all sites (53). The relative substitution rate for a variable position, $v > 1$, indicates a substitution rate higher than that averaged for all sites in the rRNA gene analyzed, while a conserved position had a relative substitution rate of $v < 1$; uncommon sites are positions occupied in <25% of organisms due to insertions. The expected variability for certain classes of positions was calculated from the consensus models. Differences between expected and observed variability were analyzed by chi-square analysis, considered significant at $P < 0.05$.

RESULTS

16S rRNA gene data set. In total, 883 complete prokaryotic genomes were available for analysis, 61 from *Archaea* and 822 from *Bacteria*. The 883 genomes represented 25 phyla or 568 unique species (see Table S1 in the supplemental material). *Proteobacteria* was the most abundant phylum (264 species), followed by *Firmicutes* (106 species), *Actinobacteria* (47 species), *Euryarchaeota* (34 species), *Crenarchaeota* (18 species), *Bacteriodetes* (13 species), *Cyanobacteria* (13 species), and *Spirochaetes* (13 species). The remaining 17 phyla were represented by only 60 species. Genome sequences from 143 species contained only a single 16S rRNA gene. The remaining 425 unique (19 archaea, 406 bacteria) species containing multiple 16S rRNA genes belonged to 18 different phyla (Table 1). Once again, the most abundant was *Proteobacteria* (210 species), followed by *Firmicutes* (98 species), *Actinobacteria* (36 species), *Euryarchaeota* (19 species), *Cyanobacteria* (11 species), and *Bacteriodetes* (10 species). The remaining 12 phyla were represented by only 41 species.

Diversity of 16S rRNA genes. The 425 unique species contained 2 to 15 copies of 16S rRNA genes per genome (2.22 ± 0.81). Intragenomic diversity of 16S rRNA genes ranged between 0 and 20.38% (0.30% ± 1.12%). Sequence diversity among the 16S rRNA genes in a genome was found in 235 species, ranging from 0.06% to 20.38% (0.55% ± 1.46%). Using the 1%-to-1.3% threshold for operational identification of species based on 16S rRNA genes (45), 24 species were found to have intragenomic diversity equal to or higher than the operational threshold (Table 2), of which the diversity was

TABLE 1. Distribution of unique prokaryotic species with more than one 16S rRNA gene by phylum

Phylum	No. of species	No. of 16S rRNA genes per genome				% diversity			
		Minimum	Avg	Maximum	σ	Minimum	Avg	Maximum	σ
Proteobacteria	210	2	4.47	15	2.54	0.00	0.20	2.09	0.34
Firmicutes	98	2	5.97	15	3.13	0.00	0.46	6.70	1.00
Actinobacteria	36	2	3.64	6	1.40	0.00	0.10	1.30	0.23
Euryarchaeota	19	2	2.74	4	0.73	0.00	0.38	5.63	1.28
Cyanobacteria	11	2	2.55	4	0.93	0.00	0.06	0.27	0.09
Bacteroidetes	10	2	4.90	7	1.79	0.00	0.34	1.30	0.47
Chlorobi	6	2	2.17	3	0.41	0.00	0.04	0.14	0.06
Chloroflexi	6	2	3.00	5	1.10	0.00	0.57	1.08	0.46
Spirochaetes	6	2	2.00	2	0.00	0.00	3.41[a]	20.38	8.31
Aquificae	5	2	2.20	3	0.45	0.00	0.04	0.20	0.09
Deinococcus-Thermus	4	2	2.75	3	0.50	0.00	0.30	1.06	0.51
Thermotogae	4	2	2.75	4	0.96	0.00	0.14	0.50	0.24
Chlamydiae	3	2	2.33	3	0.58	0.00	0.45	1.34	0.77
Dictyoglomi	2	2	2.00	2	0.00	0.00	0.03	0.06	0.04
Tenericutes	2	2	2.00	2	0.00	0.00	0.04	0.07	0.05
Acidobacteria	1	2	2.00	2	0.00	0.00	0.00	0.00	0.00
Fusobacteria	1	5	5.00	5	0.00	0.14	0.14	0.14	0.00
Verrucomicrobia	1	3	3.00	3	0.00	0.00	0.00	0.00	0.00
Summary	425	2	4.45	15	2.66	0.00	0.30	20.38	1.12

[a] The high average diversity is due to the 20.38% diversity in *Borrelia afzelii*.

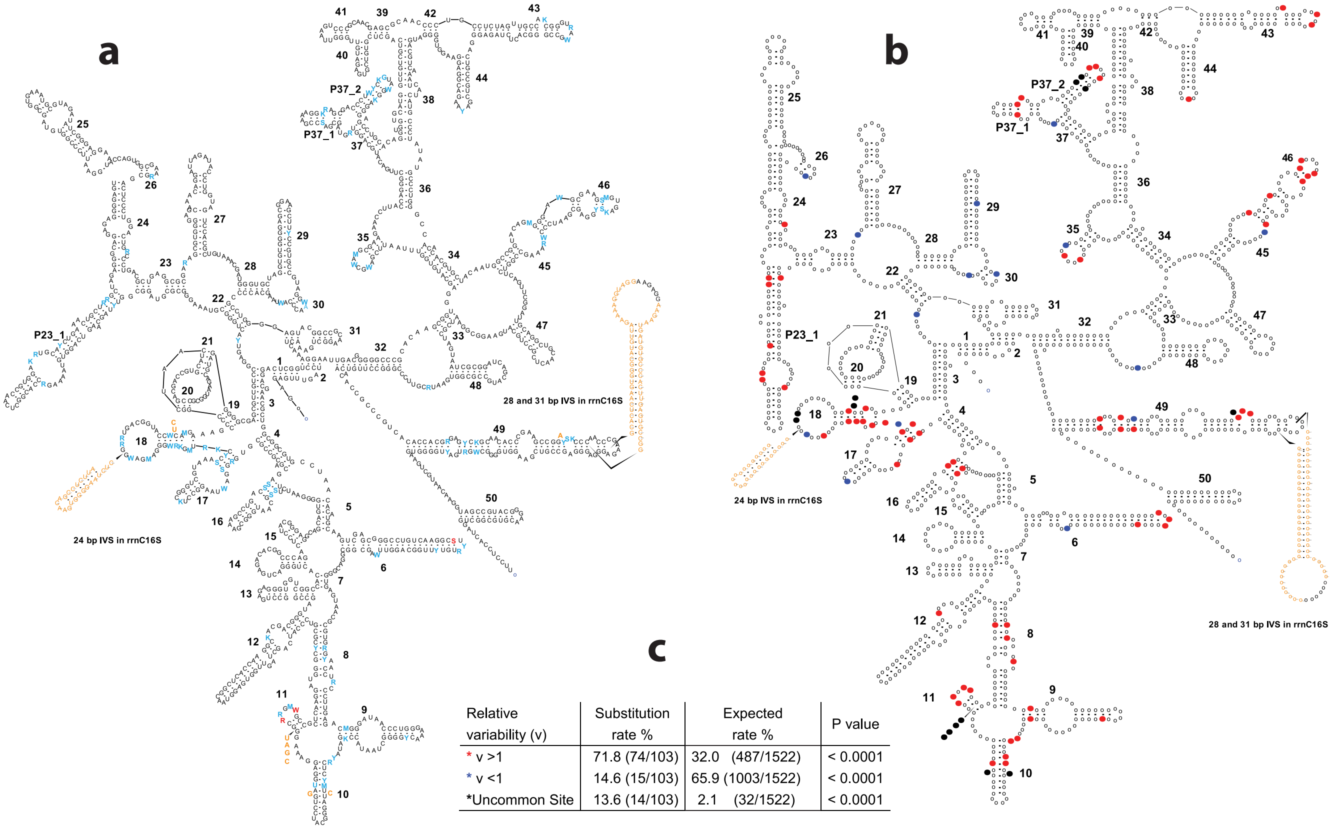
borderline (between 1% and 1.3%) in 10 species and ≥1.3% in 14 species. In particular, this phenomenon was seen in 7 bacterial species associated with the human microbiome or diseases (Table 2), including *Escherichia coli* (1.10%), *Bacillus subtilis* (1.16%), *Pseudomonas stutzeri* (1.23%), *Bacteroides thetaiotaomicron* (1.30%), *Bifidobacterium adolescentis* (1.30%), "*Candidatus* Protochlamydia amoebophila" (1.34%), and *Borrelia afzelii* (20.38%) (15, 19, 20, 22, 39, 46, 55). The sequence diversity observed in these species appears to follow five patterns: intervening sequences, regional diversity, random diversity, pseudogene, and gene truncation/partial rRNA operon.

Species with IVS. IVS in 16S rRNA genes are associated with significant intragenomic diversity among 16S rRNA genes in four species. The diversity seems related to regions near

TABLE 2. Species with significant diversity among paralogous 16S rRNA genes

Species	No. of copies	% diversity	Patterns	% diversity (masked)
Deinococcus geothermalis	3	1.06	Random	1.09
Chloroflexus sp. strain Y-400-fl	3	1.08	Regional	0.31
Shewanella sp. strain ANA-3	9	1.09	Partial operon, regional	0.16
Escherichia coli[a]	7	1.10	Partial operon, regional	0.31
Carboxydothermus hydrogenoformans	4	1.14	Regional	1.24
Bacillus clausii	7	1.15	Partial operon, regional	0.08
Bacillus subtilis[a]	10	1.16	Random	0.39
Desulfotalea psychrophila	7	1.16	Regional	0.00
Geobacillus thermodenitrificans	10	1.22	Regional	1.09
Pseudomonas stutzeri[a]	4	1.23	Regional	0.47
Bacteroides thetaiotaomicron[a]	5	1.30	Regional	0.66
Bifidobacterium adolescentis[a]	5	1.30	Partial operon, regional	0.39
"*Candidatus* Protochlamydia amoebophila"[a]	3	1.34	Gene truncation, regional	1.53
Caldicellulosiruptor saccharolyticus	3	1.35	Regional	0.31
Shewanella baltica	10	1.36	Partial operon, regional	0.08
Shewanella woodyi	10	1.56	Regional	0.08
Heliobacterium modesticaldum	8	1.58	Regional	0.23
Syntrophomonas wolfei	3	1.67	Partial operon, IVS	0.16
Photobacterium profundum	14	1.76	IVS	0.08
Desulfitobacterium hafniense	5	1.80	IVS	0.08
Clostridium cellulolyticum	8	2.07	Regional	0.32
Haloarcula marismortui	2	5.63	Regional	4.86
Thermoanaerobacter tengcongensis	4	6.70	IVS	5.01
Borrelia afzelii[a]	2	20.38	Pseudogene	NA[b]
Avg	6.08	2.55		

[a] Species found in humans.
[b] NA, not alignable.



Relative variability (v)	Substitution rate %	Expected rate %	P value
* v >1	71.8 (74/103)	32.0 (487/1522)	< 0.0001
* v <1	14.6 (15/103)	65.9 (1003/1522)	< 0.0001
*Uncommon Site	13.6 (14/103)	2.1 (32/1522)	< 0.0001

FIG. 1. Secondary structures of 16S rRNA genes from *Thermoanaerobacter tengcongensis*. (a) The distribution of substitutions is shown on the secondary structure predicted for *rrnB16S*, based on free energy minimization (31) with the consensus 16S rRNA gene model as reference (53). Positions that differ between *rrnB16S* and *rrnC16S* are shown in colored letters. Conservative changes located in loops or compensatory changes due to covariation in stems are shown in blue; changes that result in alteration of secondary structures are shown in red. Insertions/deletions are shown in brown. Substitutions are coded as follows: K = G or U, M = A or C, R = A or G, S = C or G, W = A or U, and Y = C or U. (b) The distribution of substitutions also is shown on the variability map based on the 2° structure models for *Thermus thermophilus* (53). Mismatched positions between *rrnB16S* and *rrnC16S* are highlighted in colors according to the position-specific relative variability rate, calculated from the consensus *rrn16S* model based on an alignment of 3,407 bacterial 16S rRNA genes (53). A position with a relative substitution rate of $v > 1$ (red) implies that it has a substitution rate higher than the average substitution rate of all its sites in the rRNA gene analyzed, while $v < 1$ (blue) indicates that the rate is lower than the average rate. Uncommon sites are positions that are occupied in <25% of organisms because of insertions, which are shown by black dots. The expected variability was calculated from the consensus models (c).

IVS. *T. tengcongensis* contains four 16S rRNA genes that can be classified into two major 16S rRNA gene alleles. *rrnA16S*, *rrnB16S*, and *rrnD16S* were nearly identical. Their diversity in relation to *rrnC16S* was 11.5% (186 of 1,620 positions), due to 3 inserts (24 nucleotides [nt], 28 nt, and 31 nt) and 103 mismatched positions (Fig. 1a). Considering only the 103 mismatched positions, the difference between the 2 rRNA alleles was 6.70%. Homologs of the 28- and 31-nt inserts also were present in *rrn16S* of *Thermoanaerobacter kivui* (96.6%) and *Thermoanaerobacter siderophilus* (89.8%), respectively, suggesting that such insertions may be widespread in *Thermoanaerobacter* (4) (Fig. 2). *Syntrophomonas wolfei* contains three 16S rRNA genes that varied in length (1,634 nt, 1,696 nt, and 1,646 nt, respectively). These three genes each contain a different IVS. *rrnA* has a 66-nt IVS starting at the 104th nt that shares no homology with the IVS found in *rrnB* and *rrnC*. The IVS in *rrnB* and *rrnC* are 130 nt and 79 nt, respectively, and each begins at the 103rd nt compared to *Escherichia coli*. After the IVS are removed, the greatest divergence is found between *rrnA* and *rrnC*, with 26 mismatches resulting in 1.67% diversity. The majority of the intragenomic difference is located in the 33-nt region directly adjacent to the 5′ end of the IVS. There are 22 mismatches in this area resulting in regional alteration and a divergence of 66.67%. *Desulfitobacterium hafniense* con-

```
T. kivui (DSM 2030)          GCACTCAAGTGGGCATGTTAGGAAAGGTAGGAAGAGGAAGAATGATGTGTCCACTTGAGTGCCGG
                             ||||||||||||||||||||| |||||||||||||||||||||||||||| |||||||||||||||
T. tengcongensis (MB4T)      GCACTCAAGTGGGCATGTTAGAAAAGGTAGGAAGAGGAAGAATGATGTGCCCACTTGAGTGCCGG
                             ||||||||||||||||| |||| ||||||| |||||| ||||||| ||||| |||||||||||||
T. siderophilus (SR4)        GCACTCAAGTGGGCATGATAGAGAAGGTAGTAAGAGGGAGAATGAAGTGCCTACTTGAGTGCCGG
```

FIG. 2. Homologs of the 28- and 31-bp inserts in *T. tengcongensis rrnC* 16S rRNA genes also are present in 16S rRNA genes of *T. kivui* and *T. siderophilus*. The 24- and 28-bp inserts were separated by a 6-nucleotide conserved sequence shown in red.

tains five 16S rRNA genes. 16S *rrnA*, 16S *rrnB*, and 16S *rrnC* are 1,554 nt long and only differed by 0.51%, but 16S *rrnD* (1,661 nt) and 16S *rrnE* (1,671 nt) each contain an IVS of 104 nt and 117 nt, respectively, inserted after the 91st nt. While the IVS are in nearly identical areas, they differ significantly (58.87%). Excluding the IVS, the maximal difference is 1.8% between 16S *rrnB* and 16S *rrnF*. The majority of the mismatches occurred within the 33-nt segment upstream of the IVS. This area of variability might have been caused by the insertion of IVS. The source for these IVS is unknown, but they appear to be rRNA in nature because they exhibit a highly complex 2° structure like rRNA. They were also found in two closely related species (16S *rrnE* IVS in *Desulfitobacterium frappieri* and 16S *rrnD* IVS in *D. frappieri* and *Desulfitobacterium dehalogenans*). *Photobacterium profundum* has 14 copies of 16S rRNA genes. There are 3 IVS of 14 nt (IVS-A), 14 nt (IVS-B), and 16 nt (IVS-C), respectively. Three genes (*rrnC*, *rrnJ*, and *rrnK*) contain no insert. Seven genes (*rrnA*, *rrnF*, *rrnL*, *rrnM*, *rrnN*, *rrnO*, and *rrnP*) contain only IVS-A. Three genes (*rrnB*, *rrnD*, and *rrnG*) contain IVS-A and IVS-C. There is only one gene (*rrnB*) that contains all three IVS. Excluding the IVS, the maximum difference was 2.09% between 16S *rrnC* and 16S *rrnE*.

Species with random or unusual regional diversity. No apparent clustering of substitutions in 16S rRNA genes was observed in *Deinococcus geothermalis* (1.06%) and *Bacillus subtilis* (1.16%).

Unusual regional diversity is associated with high overall diversity among paralogous 16S rRNA genes in four species. *Shewanella woodyi* has 10 16S rRNA genes. The greatest divergence of 1.56% occurs between 16S *rrnC* and 16S *rrnD*. 16S *rrnC* has two regions in which the majority of the variability lies. The first spans 41 nt between positions 995 and 1035 in variable region 6 (V6) (7) and shows 13 substitutions, which result in a 31.7% regional divergence. The second region is only 11 bp long between positions 1128 and 1138 (V7), with six substitutions for a regional divergence of 54.5%. *Heliobacterium modesticaldum* has 10 16S rRNA genes, nine of which are 1,510 nt and one of which is 1,516 nt in length. Maximal diversity is 1.58%, between 16S *rrnF* and *rrnC*. This diversity is in part due to a 20-nt region found in V2 (170th to 189th nt) that exhibits 13 mismatches for a 65% regional divergence. In addition, 16S *rrnF* has a 6-nt insert between nt 1490 and 1495 (V9). *Clostridium cellulolyticum* has eight 16S rRNA genes. The maximal divergence is 2.07%, between 16S *rrnE* and *rrnA*. This high divergence is found mainly in a 123-nt segment spanning V1 and V2 (nt 82 to 204). Within this segment in 16S *rrnE*, there are 22 substitutions and 3 indels of 1 base each, compared with 16S *rrnA*. The segment ends with a 5-nt deletion, and the rest of the 3′ strand is rather homologous to the other 7 copies. The variable area shows a divergence of 22.76% compared with the same area in *rrnA* (28 mismatches among 123 nt). The genome of *H. marismortui* has three 16S rRNA genes on two chromosomes. Although *rrnA* and *rrnC* are virtually identical, *rrnB* differs substantially from *rrnA* (5.63%), due to 83 substitutions that are mainly concentrated in three regions. Region 1 (positions 59 to 184; V1 to V2) has 11 substitutions for a regional difference of 8.7%. Region 2 (positions 524 to 840; V4 to V5) has 52 substitutions for a difference of 16.4%. Region 3 (positions 1054 to 1133, involving the region between V6 and V7, as well as portion of V7) has 10 substitutions for a difference of 16.67%.

Species with nonfunctional rRNA gene. *Borrelia afzelii* contains two 16S rRNA genes found on two separate chromosomes that differ by 20.38%. The lengths of *B. afzelii rrnA* and *rrnB* are 1,509 nt and 1,537 nt, respectively, when compared to *E. coli* as a template. While *rrnB* is closely related to 16S rRNA genes in other species of *Borrelia*, *rrnA* is significantly different. Alignment of the two genes results in a 1,570-nt overlap, of which 320 bp are mismatched (225 substitutions and 95 indels). These mismatches do not follow any type of aforementioned patterns and are distributed throughout the alignment evenly. This level of random diversity may have occurred due to the escape from constraint in a nonfunctional 16S rRNA pseudogene. There has not been any experiment designed to examine the expression of *rrnA* of *B. afzelii*. It does not appear to be horizontally transferred into *B afzelii* from other species, as it is closer to *rrnB* of *B. afzelii* than to a species in any other genus.

Species with truncation of 16S rRNA genes or partial rRNA operons. Truncation of 16S rRNA genes or partial rRNA operons are a form of structural diversity among 16S rRNA genes in a genome. Truncated 16S rRNA genes were only found in two species. "*Candidatus* Protochlamydia amoebophila" contains three 16S rRNA genes with 16S *rrn*A (1,481 nt) and 16S *rrnC* (1,481 nt) being identical, but the *rrnB* operon contains a truncated 23S gene as well as a truncated 16S gene, although the 5S gene is intact. The 16S gene was missing its 3′ end and was fused with the 23S gene on its 5′ end. The truncated 16S *rrnB* was 898 nt long and had 12 mismatches (8 substitutions and 4 indels) resulting in a diversity of 1.34%. When the missing 3′ portion was searched for in the genome, only two genes were found belonging to *rrnA* and *rrnC*, indicating that 16S *rrnB* truly is truncated. *Bacillus amyloliquefaciens* FZB42 has 10 16S rRNA genes, all of which are nearly identical (<0.5% diversity). 16S *rrnF* is only 599 nt in length, corresponding to the 3′ portion of the 16S rRNA gene. Again a search of the whole genome of *B. amyloliquefaciens* for the 5′ portion of the 16S rRNA gene found only nine copies that are associated with the nine complete 16S genes.

The absence of a whole 16S rRNA gene in a rRNA operon, as evidenced by the presence of 23S or 5S rRNA genes but absence of 16S rRNA gene, was observed in rRNA operons in 95 species (see Table S2 in the supplemental material). The 23S or 5S rRNA genes in the partial rRNA operon appear functional because none of the genes exhibit excessive random mutations characteristic of a pseudogene. Interestingly, intragenomic diversity among 16S rRNA genes in 6 of the 95 species was borderline or slightly above the 1 to 1.3% threshold for separation of species (Table 2). These species include *Shewanella* sp. strain ANA-3 (1.09%), *Escherichia coli* (1.10%), *Bacillus clausii* (1.15%), *Bifidobacterium adolescentis* (1.30%), *Shewanella baltica* (1.36%), and *Syntrophomonas wolfei* (1.67%). As described before, the high diversity in *S. wolfei* was also associated with IVS in 16S rRNA genes.

Conservation of secondary structures. In 18 of the 24 species with significant intragenomic diversity (Table 2), the majority of substitutions in each species were conserved at the 2° structure level (Table 3), and the diversity at the secondary structure level is <1%. For example, at the 2° structure level,

TABLE 3. Effect of diversity on the conservation of 2° structure

Organism	Diversity		No. of substitutions	No. of conservative changes				No. of nonconservative changes			
	1°	2°		Covariation	G:U = G · C	In loop	Indel	Stem-loop transition	Indel	Regional complex rearrangement	Regional alteration of 2° structure
Deinococcus geothermalis	1.06	0.20	16	0	6	7	0	3	0	0	0
Chloroflexus sp. strain Y-400-fl	1.08	0.41	16	2	1	6	1	6	0	0	0
Shewanella sp. strain ANA-3	1.09	0.13	17	6	6	3	0	2	0	0	0
Escherichia coli[a]	1.10	0.32	17	2	3	6	1	4	1	0	0
Carboxydothermus hydrogenoformans	1.14	0.06	18	2	3	11	1	1	0	0	0
Bacillus clausii	1.15	0.45	18	6	2	3	0	5	2	0	0
Bacillus subtilis[a]	1.16	0.58	18	0	3	6	0	4	5	0	0
Desulfotalea psychrophila	1.16	0.52	18	6	0	1	3	8	0	0	0
Geobacillus thermodenitrificans	1.22	0.32	19	0	1	8	5	5	0	0	0
Pseudomonas stutzeri[a]	1.23	0.06	19	8	5	5	0	1	0	0	0
Bacteroides thetaiotaomicron[a]	1.30	0.07	18	4	1	12	0	0	1	0	0
Bifidobacterium adolescentis[a]	1.30	0.26	20	8	3	1	4	3	1	0	0
"*Candidatus* Protochlamydia amoebophila"[a]	1.34	1.34	12	0	0	0	0	0	0	0	12
Caldicellulosiruptor saccharolyticus	1.35	0.00	21	8	5	7	1	0	0	0	0
Shewanella baltica	1.36	0.19	21	10	4	4	0	3	0	0	0
Shewanella woodyi	1.56	0.98	24	0	4	2	3	2	0	13	0
Heliobacterium modesticaldum	1.58	0.60	24	12	2	1	2	7	0	0	0
Syntrophomonas wolfei	1.66	1.46	26	0	1	2	0	1	0	0	22
Desulfitobacterium hafniense	1.80	1.16	28	6	3	1	0	1	0	0	17
Photobacterium profundum	2.09	0.98	32	8	3	6	0	3	0	12	0
Clostridium cellulolyticum	2.13	1.76	35	4	1	1	0	2	0	27	0
Haloarcula marismortui	5.63	0.82	83	51	9	11	0	12	0	0	0
Thermoanaerobacter tengcongensis	6.70	0.52	103	34	10	57	4	8	0	0	0
Borrelia afzelii PKo	20.38	6.88	320	16	82	74	40	53	55	0	0

[a] Species found in humans.

95 (92.2%) of the 103 mismatches between *rrnB16S* and *rrnC16S* in *T. tengcongensis* are conserved (Fig. 1a), composed of 48 mismatches in loops, 36 covariations, 4 insertions, and 7 GU:GC conversions. Thus, only 8 (0.5%) (3 substitutions and 5 insertions) of the 1,527 total positions are predicted to alter 2° structure, causing a stem-loop transition. Superimposition of the mismatches onto the variability map of the *Thermus thermophilus* 16S rRNA gene (31) (Fig. 1b) revealed that 74 (71.8%) of the 103 mismatch positions occurred at highly variable positions, significantly higher than expected by chance (32.0%; $P < 0.0001$) (Fig. 1c). The 29 mismatches at conserved positions account for 1.9% of the 1,527 positions in *rrn16S*. The highly conserved 2° structure indicates constraints from ribosomal proteins, indicating that *rrnC16S* is functional. In the 2° structure of *H. marismortui*, 71 (85.5%) of the 83 mismatches were conserved. Only 12 (0.82% of the 1,472 total positions) are predicted to cause an alteration, in the form of stem-loop transitions. Similarly, for *H. modesticaldum*, 17 (70.83%) of the 24 mismatches are conserved, while only 7 mismatches (0.46% of the 1,516 total positions) are predicted to alter the 2° structure, causing stem-loop transitions. In the remaining 16 species, there are total of 498 substitutions between the 16 pairs of most diversified 16S rRNA genes. The large majority (411; 82.5%) of the substitutions do not cause alteration of the 2° structure.

In the remaining 6 of the 24 species, either the majority of the substitutions are not conserved at the 2° structure level or the diversity at the 2° structure level remains >1%, of which 4 species are associated with exceptional regional diversity and 2 with nonfunctional 16S rRNA genes.

In two species, thermodynamic folding showed that substitutions caused complex regional rearrangement of base pairing, but the resultant 2° structures still topographically retained the main features of the original ones (Fig. 3 and Table 3), similar to the complex rearrangement described in 23S rRNA genes (37). In *S. woodyi*, 9 of the 24 mismatches between *rrnC16S* and *rrnD16S* were conserved, while the other 15 mismatches (0.98% of the 1,537 total positions) are predicted to alter the 2º structure. The alteration is due to 2 substitutions causing stem-loop transitions and 13 substitutions resulting in regional complex rearrangement (Fig. 3a and b). Likewise, in *C. cellulolyticum*, 28 of the 35 mismatches between *rrnA16S* and *rrnE16S* (1.70% of the 1,645 total positions) are predicted to alter the 2° structure, of which 28 are involved in regional complex rearrangement (Fig. 3c and d). Besides these two species, *P. profundum* also exhibits features of complex regional rearrangement, although the majority of substitutions (17/32) in this species are conserved (Table 3). The remaining 15 (0.98% of 1,533 total positions) are associated with alteration of the 2° structures, of which 12 are concentrated in an area of regional complex rearrangement (Fig. 3e and f). In the other two species, substitutions directly caused regional alteration in the 2° structures without complex rearrangement (Table 3). The altered regions are located directly upstream of IVS insertion sites. In *D. hafniense*, there were 28 total substitutions, of which 18 (1.16% of the 1,555 total positions) caused

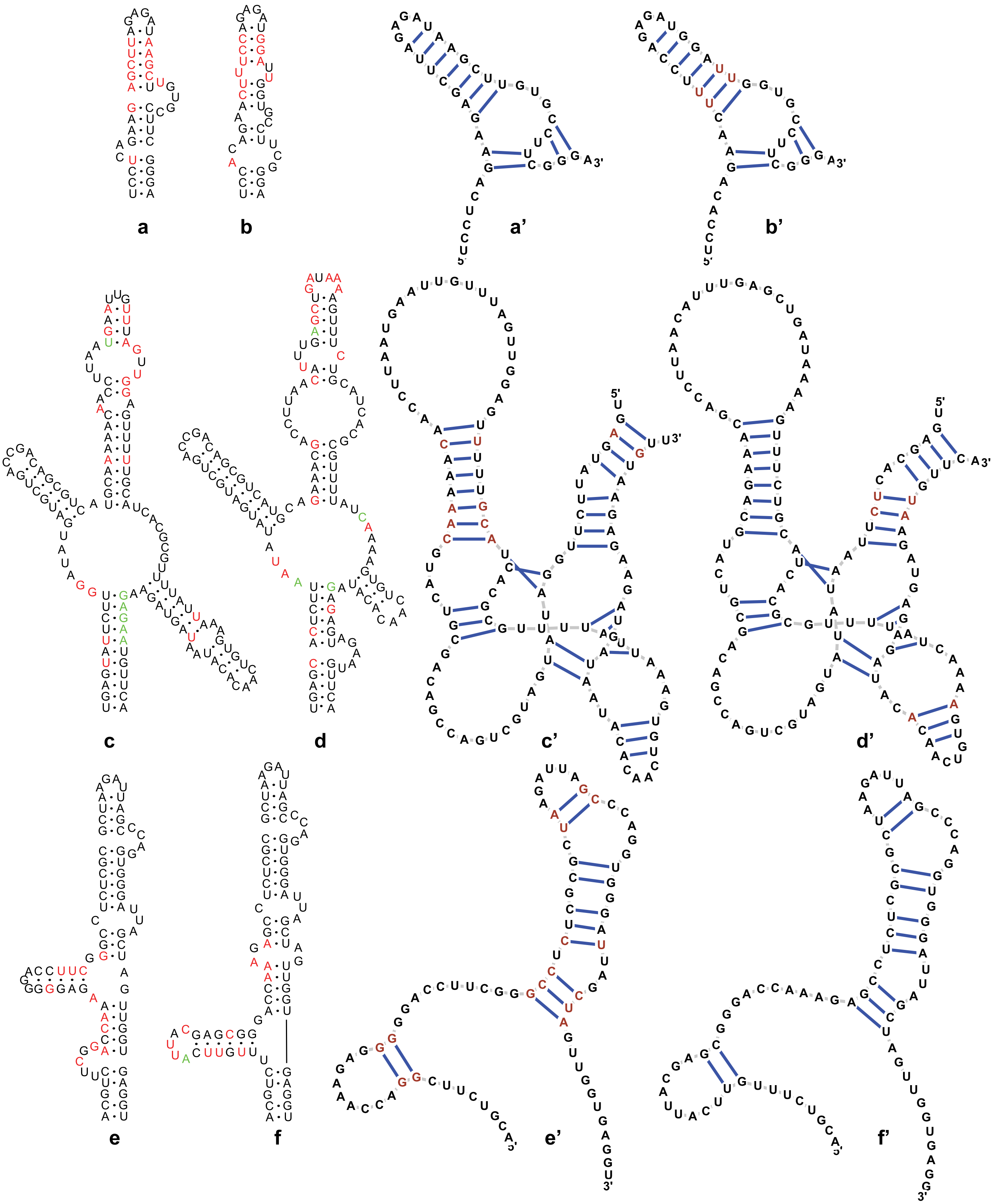


FIG. 3. Conservation of 2° structure by complex rearrangement of base pairing and substitutions in 16S rRNA genes of *S. woodyi* (a and a′ and b and b′), *C. cellulolyticum* (c and c′ and d and d′), and *P. profundum* (e and e′ and f and f′). Each molecule was folded using a program based on thermodynamics (a, b, c, d, e, and f), as well as a program based on multiple sequence alignment (a′, b′, c′, d′, e′, and f′), as described in Materials and Methods. For thermodynamic folding, the regions 50 bp upstream of the first mutation and 50 bp downstream of the last mutation were used to create the structures for each rRNA molecule, while only the area of interest is shown. Nucleotides related to substitutions are

an alteration. One caused stem-loop transition, while 17 were found in an area of regional alteration directly upstream of the IVS insertion site. In *S. wolfei*, there were 26 total mismatches between *rrnA16S* and *rrnC16S*: 22 resulting in regional alteration of the 2° structure immediately upstream of the IVS and 1 resulting in a stem-loop transition (1.46% of the 1,570 total positions). For these 16S rRNA genes that displayed high levels of regional diversity, the regions in question were also folded using the multiple-sequence-alignment approach (5). As shown in Fig. 3 (a′ to f′), a consensus 2° structure for the pair of diversified regions can be constructed if noncanonical base pairing is allowed (35, 36), which provides another possibility for conservation of the 2° structure among highly diversified regions.

The truncated 16S rRNA gene (16S *rrnB*) in "*Candidatus* Protochlamydia amoebophila" is probably a nonfunctional gene. It is only 898 nt long and contains mismatches all associated with alteration of the 2° structure.

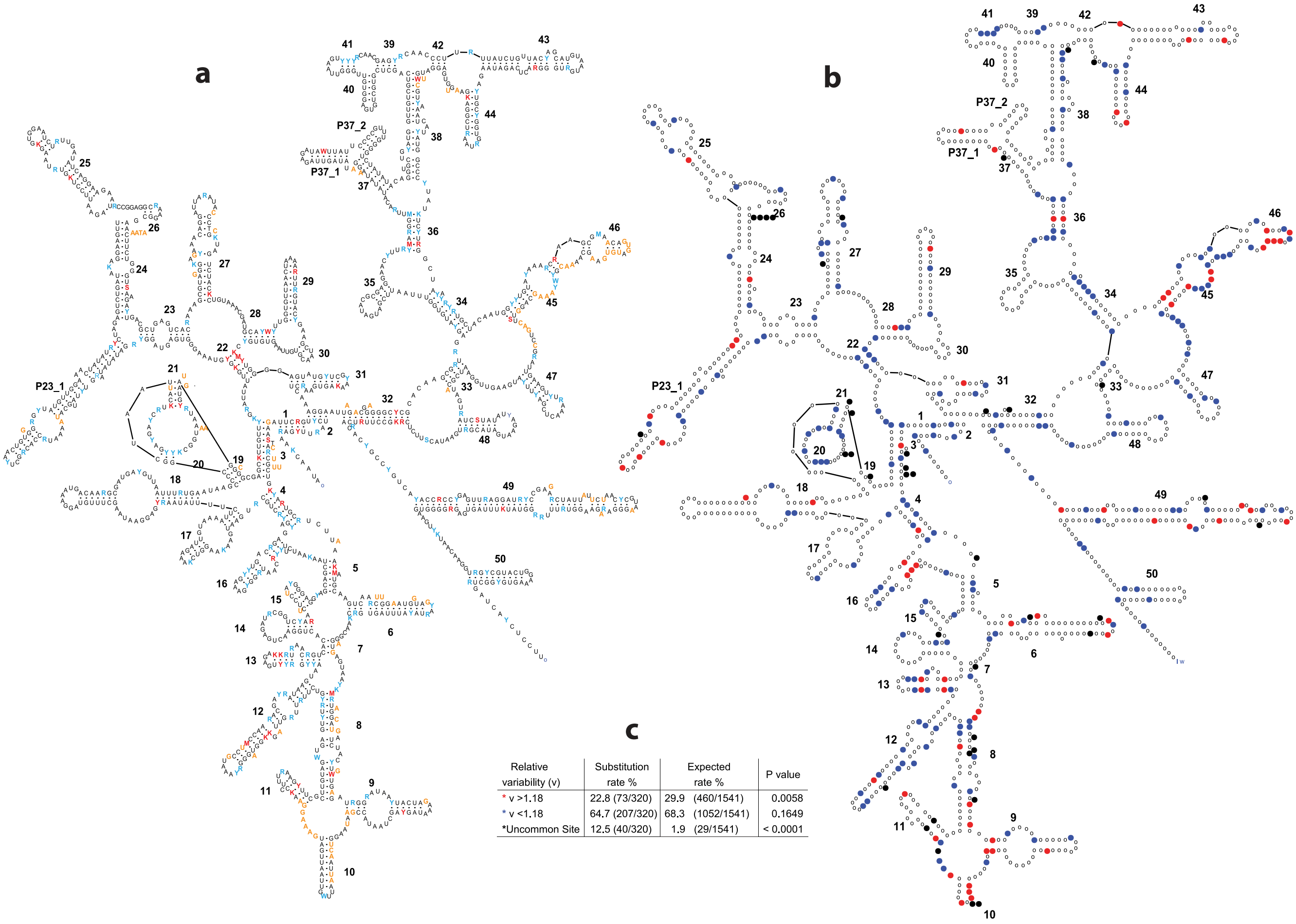
Borrelia afzelii is unique among the genomes deciphered to date, in that the divergence remains extremely high (6.68%) even after examination of the 2° structure of the genes. The 320 substitutions between *rrnA* (1,509 nt) and *rrnB* (1,537 nt) resulted in 108 positions (6.88%) of 2° structure alteration (55 indels and 53 stem-loop transitions) over a 1,570-nt alignment (Fig. 4a). Superimposition of the mismatches to the variability map of *E. coli* 16S rRNA gene (31) (Fig. 4b) revealed that only 73 (22.8%) of the 320 substitutions occurred at highly variable positions, significantly less than expected by chance (29.9%; $P < 0.0058$) (Fig. 4c). Similarly, the substitution rate at conserved positions (207/320; 64.7%) was also less than expected by chance (68.3%) but not statistically different ($P < 0.1649$) (Fig. 4c). Conversely, a large number of the substitutions (40; 12.5%) were located at sites not commonly seen the majority of 16S rRNA genes (Fig. 4c).

Effect of masking hypervariable positions. Normally, for phylogenetic analysis, only those positions that can be well aligned between most sequences are included in the analysis. Hypervariable regions are usually "masked out." To assess the effect of masking hypervariable regions on intragenomic variation of 16S rRNA genes, we aligned 16S rRNA genes from the 24 highly diversified species (Table 2) and masked the aligned sequences with Lane mask (12, 27). Intragenomic differences were recalculated on the masked sequences, and the results are shown in Table 2. The effect of masking hypervariable positions was remarkable for the 14 of the 24 species with intragenomic diversity of 16S rRNA genes greater than 1%. The masking reduced the diversity from between 1.06% and 2.07% to <0.66% (Table 2). This level of diversity will not have a significant impact on phylogenetic inference. However, the variation after masking remained high for *H. marismortui* (4.86%) and *T. tengcongensis* (5.01%), "*Candidatus* Protochlamydia amoebophila" (1.53%), *Carboxydothermus hydrogenoformans* (1.24%), *Deinococcus geothermalis* (1.09%), and *Geobacillus thermodenitrificans* (1.09%). For *B. afzelii*, the two 16S rRNA genes are too diversified to be aligned using available algorithms (the original alignment of full sequences of these two genes was done manually). Such a high level of diversity after masking will be troublesome not only for threshold-based taxonomic assignment using full-length sequences, but also for phylogenetic inference using masked sequences because 16S rRNA genes from within the same genome are not monophyletic in these species.

DISCUSSION

We analyzed 16S rRNA genes from genomes representing 425 prokaryotic species to examine the evidence supporting the theory of ribosomal constraint of rRNA structures at the intragenomic level. Our findings support the hypothesis that individual rRNA genes within a genome are conserved due to such structural constraints. Prokaryotes comply with the constraint at both the primary and secondary structure levels. In organisms containing multiple rRNA genes, the homogeneity of primary sequences is believed to be maintained through gene conversion by homologous recombination (18), as a form of concerted evolution (1). Although there are no criteria to define the ribosomal constraint, observations suggest that 1 to 1.3% diversity between 16S rRNA genes is tolerated by strains from the same species (44). Our data indicate that homologous recombination is highly effective in maintenance of 16S rRNA homogeneity, as the diversity is confined within 1% in all but only 24 of the 425 species. The exceptionally high diversity in the 24 species does not violate the ribosomal constraints, as they can be explained at the 2° structure level, the ultimate line of ribosomal constraint (56). For example, in *T. tengcongensis*, the 2° structure is well conserved despite 6.70% diversity in the primary structure (Table 3). Similarly, in *H. marismortui*, 71 of the 83 (85.5%) substitutions in the primary structure are conserved at the 2° structure level. The type of diversity in *B. afzelii* represents a novel example to support the theory of ribosomal constraint. The two 16S rRNA genes in this species differ by 20.38% in the 1° structure. Unlike the pattern seen in *T. tengcongensis* and *H. marismortui*, substitutions in *B. afzelii* occurred in a nonselective fashion among both the conserved and variable positions, resulting in drastic alteration of the 2° structure (6.88%) and a large number of indels ($n = 95$). This level of random diversity in *B. afzelii* is not violating the concept of ribosomal constraint; rather, it may have occurred due to the escape from constraint in a nonfunctional 16S rRNA pseudogene. Similarly, the truncated 16S rRNA gene (16S *rrnB*) with high diversity at the 2° structural level in "*Candidatus* Protochlamydia amoebophila" is probably a nonfunctional gene.

highlighted in red and indels in green. For folding using the multiple-sequence-alignment approach, nucleotides making up noncanonical base pairs are highlighted in brown (35, 36). Segments of *rrn16S* of *S. woodyi* correspond to positions 965 through 1065 of *rrnC16S* and *rrnD16S*. The segments of *rrnC16S* (a and a′) and *rrnD16S* (b and b′) differ by 13 positions, all substitutions. Segments of *rrn16S* of *C. cellulolyticum* correspond to positions 31 through 256 of *rrnA16S* and *rrnE16S*. The segments of *rrnA16S* (c and c′) and *rrnE16S* (d and d′) differ by 28 positions, including 10 indels and 18 substitutions. Segments of *rrn16S* of *P. profundum* correspond to positions 129 through 254 of *rrnC16S* and *rrnE16S*. The segments of *rrnC16S* (e and e′) and *rrnE16S* (f and f′) differ by 12 positions, including 1 indel and 11 substitutions.



Relative variability (v)	Substitution rate %	Expected rate %	P value
* v >1.18	22.8 (73/320)	29.9 (460/1541)	0.0058
* v <1.18	64.7 (207/320)	68.3 (1052/1541)	0.1649
*Uncommon Site	12.5 (40/320)	1.9 (29/1541)	< 0.0001

FIG. 4. Secondary structures of 16S rRNA genes from *Borrelia afzelii*. (a) The distribution of substitutions is shown on the secondary structure predicted for *rrnB16S*, based on free energy minimization (31) with the consensus 16S rRNA gene model as the reference (53). Positions that differ between *rrnB16S* and *rrnA16S* are shown in colored letters. Conservative changes located in loops or compensatory changes due to covariation in stems are shown in blue; changes that result in alteration of secondary structures are shown in red. Insertions/deletions are shown in brown. Substitutions are coded as follows: K = G or U, M = A or C, R = A or G, S = C or G, W = A or U, and Y = C or U. (b) The distribution of substitutions is also shown on the variability map based on the 2° structure models for *Escherichia coli* (53). Substitutions between *rrnB16S* and *rrnA16S* are highlighted in colors according to the position-specific relative variability rate calculated from the consensus *rrn16S* model based on an alignment of 3,407 bacterial 16S rRNA genes (53). A position with a relative substitution rate of $v > 1.18$ (red) implies that it has a substitution rate higher than the average substitution rate of all of its sites in the rRNA gene analyzed, while $v < 1.18$ (blue) indicates that the rate is lower than the average. Uncommon sites are positions that are occupied in <25% of organisms because of insertions, which are shown by black dots. The expected variability was calculated from the consensus models (c).

For the four species with exceptionally high regional diversity (Table 3), the classical thermodynamic folding failed to show a significant effect of ribosomal constraint at the 2° structural level, but consensus 2° structure can be obtained for the pair of diversified regions if noncanonical base pairs are allowed (35, 36), providing another possibility for conservation of 2° structure among highly diversified regions. Thus, this largest survey of intragenomic diversity indicates that all prokaryotic 16S rRNA genes conform to the theory of ribosomal constraints.

Gene truncation is another form of intragenomic variation of 16S rRNA genes. It ranges from complete absence of the 16S rRNA gene in an rRNA operon to partial truncation of the 16S rRNA gene. A complete absence of a 16S rRNA gene is a common phenomenon involving rRNA operons in 95 of the 568 species (see Table S2 in the supplemental material). The absence of the 16S rRNA gene in an rRNA operon is associated with an increase in intragenomic diversity in 6 species (Table 2).

Intervening sequences (IVS) have been observed in rRNA genes previously (4, 37). IVS in 16S rRNA genes were found in 4 species in this study. IVS appear to correlate with high level of intragenomic variation among 16S rRNA genes. In one species, *T. tengcongensis*, the diversity (6.70%) in the primary structure far exceeds 1 to 1.3% (the species boundary), while in the other three species, *P. profundum*, *D. hafniense*, and *S. wolfei*, IVS are associated with above average diversity, at 2.09%, 1.8%, and 1.67%, respectively. The variations appear to be due to sequence diversity at the region 5′ upstream from the IVS, likely associated with the insertion event of the IVS. Compared with IVS found in 23S rRNA genes (37), those in 16S rRNA genes are smaller and do not contain open reading frames. The small IVS found in this study often can be folded to an rRNA-like secondary structure.

It is well known that there is wide variation of copy numbers of 16S rRNA gene among various species (28, 42). Currently, it is common practice to describe the composition of a microbial community using 16S gene composition rather than cell composition. It would be desirable to convert 16S gene composition to cell composition, but for a large number of organisms in a complex microbiome, this conversion is not possible because of the lack of knowledge about the copy numbers of 16S rRNA gene in their genomes. Here is a hypothetical example to illustrate the difference between the 16S gene composition and cell composition, in which a microbial community contains 100 bacterial cells: 90 cells from *Borrelia turicatae* and 10 cells from *Brevibacillus brevis*. Because there is one 16S rRNA gene per cell for *B. turicatae* and 15 16S rRNA genes per cell for *Brevibacillus brevis*, this community contains 240 16S rRNA genes, 90 from *B. turicatae* and 150 from *Brevibacillus brevis*. Consequently, this community is dominated by cells from *B. turicatae* (90/100) and by 16S rRNA genes from *Brevibacillus brevis* (150/240). Thus, 16S gene composition is an acceptable way to describe a microbial community with the understanding of the difference between the 16S gene composition and cell composition.

In 3 of the 24 diversified species, *H. marismortui*, *T. tengcongensis*, and *B. afzelii*, the intragenomic diversity between 16S rRNA genes correlated with differences in GC content. *H. marismortui* is a halophilic red archaeon found in the Dead Sea, a high-saline environment (16). *T. tengcongensis*, isolated from a hot spring in Tengchong, China, grows optimally at 80°C (4). It is known that these organisms thrive in such environments due to adaptations in protein structure, metabolic strategies, and physiologic responses. Data from this work support the recent observation by Lopez-Lopez et al. that the spectrum of adaptation may extend to diversification in rRNA genes (29). The diversified copies of 16S rRNA genes have a higher GC content than their regular counterpart (58.4% versus 56.7% for *H. marismortui* and 60.3% versus 59.2% for *T. tengcongensis*). In *H. marismortui*, the rRNA operon containing the high-GC 16S rRNA gene appears to have a different promoter and is preferentially expressed at higher temperature (29). Thus, diversification in rRNA genes could be selected to meet survival needs under different environmental conditions. *B. afzelii* normally resides in the digestive tract of ticks or fleas and is transmitted to humans via the insect's saliva following a bite (19). It is one of the four major *Borrelia* species causing Lyme disease. Of the two 16S rRNA genes in *B. afzelii*, the pseudogene has a much lower GC content (38.1%) than the functional copy (46.5%). It appears the random mutations in the pseudogene have been bringing its GC content toward the baseline for the whole genome (28%).

16S rRNA has been a widely used marker for taxonomic classification of prokaryotic organisms. By the operational definition, two 16S rRNA genes differing by 1 to 1.3% or more represent two different species (44). Using such thresholds would artificially classify the 24 species with intragenomic diversity of 16S rRNA genes of >1% into more than 24 species. In a population-based microbiome survey, the presence of such intragenomic sequence divergence would result in overestimation of the species diversity of the microbiome and possibly skew the community structure, if these species are prevalent in the microbiome. Besides confirming the previous observation of the presence of highly diversified 16S rRNA in single prokaryotic genomes (2), our study provides an estimate of the extent of this phenomenon. It involves both *Bacteria* and *Archaea*, ranging from extremophiles to human pathogens. The overall probability to encounter a species with significant intragenomic diversity among its 16S rRNA genes is about 4.2% (24 in 569 unique species or 24 in 425 unique species with multiple 16S rRNA genes). This phenomenon will have an impact on analysis of the human microbiome because 7 species found in the human microbiome or diseases exhibit intragenomic 16S rRNA diversity greater than the threshold for defining species.

The definition for a prokaryotic species is polyphasic, in that it requires a distinct set of biological characteristics and corresponding DNA reassociation values of >70%. However, there is not a simple universal definition. 16S rRNA genes have been used as a surrogate maker for operationally defining species. Initially, a >3% difference between 16S rRNA genes from two organisms was required to claim the two organisms belong to different species (43). Later, the threshold was lowered to 1 to 1.3% (44). This operational definition is helpful in taxonomic classification using 16S rRNA genes, especially for studies of complex microbiomes using cultivation-independent techniques in which biological characteristics and DNA reassociation values cannot be determined for individual bacterial cells/species. Nevertheless, it is critical to understand the limitations of the 16S rRNA-based operational definition for species. The main limitation is that 16S rRNA genes evolve at different rates, but the operational species threshold (1 to 1.3%) is relatively rigid. As a consequence, closely related species that evolve slowly will be grouped as a single species by the operational definition, such as *Streptococcus pseudopneumoniae* and *Streptococcus pneumoniae* (3), which differ by only 5 bp between their 16S rRNA genes (a 0.03% difference). Another limitation is that 16S does not represent the entire genomic content that determines the biological characteristics for a species. Significant differences in genome composition may be present in bacterial species that are completely identical or that differ only slightly in 16S rRNA genes. For example, isolates of *Vibrio splendidus* exhibit up to 25% genotypic difference (45), and strains of *E. coli* may differ up to 40% in the number of genes in their genomes (24, 38). The three members in the *Bacillus cereus* group (*B. cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*) can be classified as a single species by their nearly identical 16S rRNA genes, but they differ greatly by the number and type of genes they harbor due to the presence of large plasmids (41). In both *E. coli* and the *B. cereus* group, these differences confer various biological capabilities and pathogenicities. Intragenomic variation of 16S rRNA genes is another limit that can be encountered when classifying species that harbor 16S rRNA genes with diversity greater than the threshold set by the operational definition (Table 2), which will lead to overestimation of species diversity in a microbiome. Thus, it can be expected that community structures determined using the 16S rRNA-based operational species definition approximate but do not necessarily reflect the true community structures.

ACKNOWLEDGMENTS

This study was supported in part by grant UH2CA140233 from the Human Microbiome Project of the NIH Roadmap Initiative and National Cancer Institute and by grant R01AI063477 from the National Institute of Allergy and Infectious Diseases. Part of this work was carried out under the auspices of the University of California, at Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231.

We thank Meisheng Zhou and David Rosmarin for assistance in verification of sequence diversity and Lauren Valentor for editing the manuscript.

REFERENCES

1. **Abdulkarim, F., and D. Hughes.** 1996. Homologous recombination between the tuf genes of *Salmonella typhimurium*. J. Mol. Biol. **260:**506–522.
2. **Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz.** 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. J. Bacteriol. **186:**2629–2635.
3. **Arbique, J. C., C. Poyart, P. Trieu-Cuot, G. Quesne, M. da Glória, S. Carvalho, A. G. Steigerwalt, R. E. Morey, D. Jackson, R. J. Davidson, and R. R. Facklam.** 2004. Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and a description of *Streptococcus pseudopneumoniae* sp. nov. J. Clin. Microbiol. **42:**4686–4696.
4. **Bao, Q., Y. Tian, W. Li, Z. Xu, Z. Xuan, S. Hu, W. Dong, J. Yang, Y. Chen, Y. Xue, Y. Xu, X. Lai, L. Huang, X. Dong, Y. Ma, L. Ling, H. Tan, R. Chen, J. Wang, J. Yu, and H. Yang.** 2002. A complete sequence of the *T. tengcongensis* genome. Genome Res. **12:**689–700.
5. **Bindewald, E., and B. A. Shapiro.** 2006. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA **12:**342–352.
6. **Carranza, S., G. Giribet, C. Ribera, J. Baguna, and M. Riutort.** 1996. Evidence that two types of 18S rDNA coexist in the genome of *Dugesia* (*Schmidtea*) *mediterranea* (*Platyhelminthes*, *Turbellaria*, *Tricladida*). Mol. Biol. Evol. **13:**824–832.
7. **Chakravorty, S., D. Helb, M. Burday, N. Connell, and D. Alland.** 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods **69:**330–339.
8. **Cilia, V., B. Lafay, and R. Christen.** 1996. Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. Mol. Biol. Evol. **13:**451–461.
9. **Clayton, R. A., G. Sutton, P. S. Hinkle, Jr., C. Bult, and C. Fields.** 1995. Intraspecific variation in small subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int. J. Syst. Bacteriol. **45:**595–599.
10. **Coenye, T., and P. Vandamme.** 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol. Lett. **228:**45–49.
11. **De Rijk, P., and R. De Wachter.** 1997. RnaViz, a program for the visualization of RNA secondary structure. Nucleic Acids Res. **25:**4679–4684.
12. **DeSantis, T. Z., Jr., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. **34:**W394–W399.
13. **Doolittle, W. F.** 1999. Phylogenetic classification and the universal tree. Science **284:**2124–2129.
14. **Eigen, M., B. Lindemann, R. Winkler-Oswatitsch, and C. H. Clarke.** 1985. Pattern analysis of 5S rRNA. Proc. Natl. Acad. Sci. U. S. A. **82:**2437–2441.
15. **Fritsche, T. R., R. K. Gautom, S. Seyedirashti, D. L. Bergeron, and T. D. Lindquist.** 1993. Occurrence of bacterial endosymbionts in *Acanthamoeba* spp. isolated from corneal and environmental specimens and contact lenses. J. Clin. Microbiol. **31:**1122–1126.
16. **Goo, Y. A., J. Roach, G. Glusman, N. S. Baliga, K. Deutsch, M. Pan, S. Kennedy, S. DasSarma, W. V. Ng, and L. Hood.** 2004. Low-pass sequencing for microbial comparative genomics. BMC Genomics **5:**3.
17. **Gunderson, J. H., M. L. Sogin, G. Wollett, M. de la Cruz Hollingdale, V. F. de la Cruz, A. P. Waters, and T. F. McCutchan.** 1987. Structurally distinct, stage-specific ribosomes occur in Plasmodium. Science **238:**933–937.
18. **Hashimoto, J. G., B. S. Stevenson, and T. M. Schmidt.** 2003. Rates and consequences of recombination between rRNA operons. J. Bacteriol. **185:**966–972.
19. **Hengge, U. R., A. Tannapfel, S. K. Tyring, R. Erbel, G. Arendt, and T. Ruzicka.** 2003. Lyme borreliosis. Lancet Infect. Dis. **3:**489–500.
20. **Hong, H. A., R. Khaneja, N. M. Tam, A. Cazzato, S. Tan, M. Urdaci, A. Brisson, A. Gasbarrini, I. Barnes, and S. M. Cutting.** 2009. *Bacillus subtilis* isolated from the human gastrointestinal tract. Res. Microbiol. **160:**134–143.
21. **Hugenholtz, P., G. W. Tyson, R. I. Webb, A. M. Wagner, and L. L. Blackall.** 2001. Investigation of candidate division TM7, a recently recognized major lineage of the domain *Bacteria* with no known pure-culture representatives. Appl. Environ. Microbiol. **67:**411–419.
22. **Johnson, T. J., and L. K. Nolan.** 2009. Pathogenomics of the virulence plasmids of *Escherichia coli*. Microbiol. Mol. Biol. Rev. **73:**750–774.
23. **Kiryu, H., Y. Tabei, T. Kin, and K. Asai.** 2007. Murlet: a practical multiple alignment tool for structural RNA sequences. Bioinformatics **23:**1588–1598.
24. **Kudva, I. T., P. S. Evans, N. T. Perna, T. J. Barrett, F. M. Ausubel, F. R. Blattner, and S. B. Calderwood.** 2002. Strains of *Escherichia coli* O157:H7 differ primarily by insertions or deletions, not single-nucleotide polymorphisms. J. Bacteriol. **184:**1873–1879.
25. **Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118–123.
26. **Küntzel, H., M. Heidrich, and B. Piechulla.** 1981. Phylogenetic tree derived from bacterial, cytosol and organelle 5S rRNA sequences. Nucleic Acids Res. **9:**1451–1461.
27. **Lane, D. J.** 1991. 16S/23S rRNA sequencing, p. 115–174. *In* E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, Chichester, United Kingdom.
28. **Lee, Z. M. P., C. Bussema III, and T. M. Schmidt.** 2009. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res. **37:**D489–D493.
29. **López-López, A., S. Benlloch, M. Bonfá, F. Rodríguez-Valera, and A. Mira.** 2007. Intragenomic 16S rDNA divergence in *Haloarcula marismortui* is an adaptation to different temperatures. J. Mol. Evol. **65:**687–696.
30. **Mashkova, T. D., T. I. Serenkova, A. M. Mazo, T. A. Avdonina, M. Y. Timofeyeva, and L. L. Kisselev.** 1981. The primary structure of oocyte and somatic 5S rRNAs from the loach *Misgurnus fossilis*. Nucleic Acids Res. **9:**2141–2151.
31. **Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner.** 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A. **101:**7287–7292.
32. **Mevarech, M., S. Hiesch-Twizer, S. Goldman, E. Yakobson, H. Eisenberg, and P. P. Dennis.** 1989. Isolation and characterization of the rRNA gene clusters of *Halobacterium marismortui*. J. Bacteriol. **171:**3479–3485.
33. **Mylvaganam, S., and P. P. Dennis.** 1992. Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics **130:**399–410.
34. **NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer.** 2009. The NIH Human Microbiome Project. Genome Res. **19:**2317–2323.
35. **Ninio, J.** 1979. Prediction of pairing schemes in RNA molecules—loop contributions and energy of wobble and non-wobble pairs. Biochimie **61:**1133–1150.
36. **Olson, W. K., M. Esguerra, Y. Xin, and X. J. Lu.** 2009. New information content in RNA base pairing deduced from quantitative analysis of high-resolution structures. Methods **47:**177–186.
37. **Pei, A., C. W. Nossa, P. Chokshi, M. J. Blaser, L. Yang, D. M. Rosmarin, and Z. Pei.** 2009. Diversity of 23S rRNA genes within individual prokaryotic genomes. PLoS One **4:**e5437.
38. **Perna, N. T., G. Plunkett III, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:**529–533.
39. **Poirel, L., J. M. Rodríguez-Martínez, N. Al Naiemi, Y. J. Debets-Ossenkopp, and P. Nordmann.** 22 March 2010, posting date. Characterization of DIM-1, an integron-encoded metallo-β-lactamase from a *Pseudomonas stutzeri* clinical isolate in The Netherlands. Antimicrob. Agents Chemother. [Epub ahead of print.]
40. **Quince, C., A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan.** 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods **6:**639–641.
41. **Rasko, D. A., M. R. Altherr, C. S. Han, and J. Ravel.** 2005. Genomics of the *Bacillus cereus* group of organisms. FEMS Microbiol. Rev. **29:**303–329.
42. **Rastogi, R., M. Wu, I. Dasgupta, and G. E. Fox.** 2009. Visualization of ribosomal RNA operon copy number distribution. BMC Microbiol. **9:**208.
43. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. **44:**846–849.
44. **Stackebrandt, E., and J. Ebers.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today **2006:**153–155.
45. **Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel, and M. F. Polz.** 2005. Genotypic

diversity within a natural coastal bacterioplankton population. Science **307:** 1311–1313.

46. **Turroni, F., J. R. Marchesi, E. Foroni, M. Gueimonde, F. Shanahan, A. Margolles, D. van Sinderen, and M. Ventura.** 2009. Microbiomic analysis of the bifidobacterial population in the human distal gut. ISME J. **3:**745–751.
47. **Wang, Y., Z. Zhang, and N. Ramanan.** 1997. The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. J. Bacteriol. **179:**3270–3276.
48. **Wegnez, M. R., R. Monier, and H. Denis.** 1972. Sequence heterogeneity of 5S rRNA in *Xenopus laevis*. FEBS Lett. **25:**13–20.
49. **Wiese, K. C., E. Glen, and A. Vasudevan.** 2005. JViz.Rna—a Java tool for RNA secondary structure visualization. IEEE Trans. Nanobioscience **4:**212–218.
50. **Woese, C. R.** 1987. Bacterial evolution. Microb. Rev. **51:**221–271.
51. **Woese, C. R.** 1998. The universal ancestor. Proc. Natl. Acad. Sci. U. S. A. **95:**6854–6859.
52. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc. Natl. Acad. Sci. U. S. A. **87:**4576–4579.
53. **Wuyts, J., Y. Van de Peer, and R. De Wachter.** 2001. Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. Nucleic Acids Res. **29:**5017–5028.
54. **Yap, W. H., Z. Zhang, and Y. Wang.** 1999. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. **181:**5201–5209.
55. **Zocco, M. A., M. E. Ainora, G. Gasbarrini, and A. Gasbarrini.** 2007. *Bacteroides thetaiotaomicron* in the gut: molecular aspects of their interaction. Dig. Liver Dis. **39:**707–712.
56. **Zwieb, C., C. Glotz, and R. Brimacombe.** 1981. Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species. Nucleic Acids Res. **9:**3621–3640.